Texila International Journal of Academic Research
Special Edition Apr 2019

The Risk of Artificial Intelligence in Cyber Security and the Role of Humans

Article by Joseph Ogaba Oche
M.Sc., MCITP, MCSE, Certified PHP Developer
E-mail: Joseph.oche@hotmail.com

Abstract

This paper will present and analyze reported failures of artificially intelligent systems and extrapolate our analysis to future AIs. I suggest that both the frequency and the seriousness of future AI failures will steadily increase. AI Safety can be improved based on ideas developed by cybersecurity experts. For narrow AIs safety failures are at the same, moderate, level of criticality as in cybersecurity, however for general AI, failures have a fundamentally different impact. A single failure of a super intelligent system may cause a catastrophic event without a chance for recovery. The goal of cybersecurity is to reduce the number of successful attacks on the system; the goal of AI Safety is to make sure zero attacks succeed in bypassing the safety mechanisms. Unfortunately, such a level of performance is unachievable. Every security system will eventually fail; there is no such thing as a 100% secure system. Future generations may look back at our time and identify it as one of intense change. In a few short decades, we have morphed from a machine-based society to an information-based society, and as this Information Age continues to mature, society has been forced to develop a new and intimate familiarity with data-driven and algorithmic systems. Artificial agents to refer to devices and decision-making aids that rely on automated, data- driven, or algorithmic learning procedures. Such agents are becoming an intrinsic part of our regular decision-making processes. Their emergence and adoption lead to a bevy of related policy questions.

Keywords: *AI Safety, Cybersecurity, Failures, Super intelligence, Algorithms, Advanced Persistent Threats (APT).*

Introduction

Artificial Intelligence or AI is the intelligence shown by machines. When any machine becomes aware of its surroundings and does something keeping that in mind in order to achieve something. Usually the term Artificial Intelligence is used when a machine behaves like a human in activities such as problem solving or learning, which is also known as Machine Learning.

The maturation of the Information Age has forced some adaptation and evolution in our laws, regulations, and policies. But the pace and intensity of technological change has often made it difficult for the policy, regulations, and laws to keep up. As has been the case in other periods of intense change, the lag in the evolution of laws and regulations can lead to significant policy gaps.

For example, data-laden societies are currently re-evaluating acceptable personal standards of privacy. This is necessary given the growing use of ubiquitous data collection and powerful, cheaply run, and readily available algorithms. The legal standards of reason- able or acceptable privacy need renegotiation to accommodate new technologies that are being adopted at pace and scale. There is a lot at stake (Ohm, 2009; Davis and Osoba, 2016): health data privacy, consumer fairness, and even the constitutional Census mandate.

A day does not go by without a news article reporting some amazing breakthrough in artificial intelligence. In fact, progress in AI has been so steady that some futurologists, such as Ray Kurzweil, project current trends into the future and anticipate what the headlines of tomorrow will bring us. Consider some developments from the world of technology:

2004 DARPA sponsors a driverless car grand challenge. Technology developed by the participants eventually allows Google to develop a driverless automobile and modify existing transportation laws.

DOI: 10.21522/TIJAR.2014.SE.19.01.Art001
ISSN: 2520-3088

2005 Honda's ASIMO humanoid robot is able to walk as fast as a human, delivering trays to customers in a restaurant setting. The same technology is now used in military robots.

2007 Computers learned to play a perfect game of checkers, and in the process opened the door for algorithms capable of searching vast databases of information.

2011 IBM's Watson wins Jeopardy against top human champions. It is currently training to provide medical advice to doctors. It is capable of mastering any domain of knowledge.

2012 Google releases its Knowledge Graph, a semantic search knowledge base, likely to be the first step toward true artificial intelligence.

2013 Facebook releases Graph Search, a semantic search engine with intimate knowledge about Facebook's users, essentially making it impossible for us to hide anything from the intelligent algorithms. 2013 BRAIN initiative aimed at reverse engineering the human brain receives 3 billion US dollars in funding by the White House, following an earlier billion-euro European initiative to accomplish the same.

2014 Chatbot convinced 33% of the judges that it was human and by doing so passed a restricted version of a Turing Test. 2015 Single piece of general software learns to outperform human players in dozens of Atari video games.

2016 Go playing deep neural network beats world champion.

From the above examples, it is easy to see that not only is progress in AI taking place, it is accelerating as the technology feeds on itself. While the intent behind the research is usually good, any developed technology could be used for good or evil purposes.

From observing exponential progress in technology, Ray Kurzweil was able to make hundreds of detailed predictions for the near and distant future. As early as 1990 he anticipated that among other things, we will see between 2010 and 2020:

- Eyeglasses that beam images onto the users' retinas to produce virtual reality (Project Glass).
- Computers featuring "virtual assistant" programs that can help the user with various daily tasks (Siri).
- Cell phones built into clothing and able to project sounds directly into the ears of their users (E-textiles).

But his projections for a somewhat distant future are truly breathtaking and scary. Kurzweil anticipates that by the year:

2029 Computers will routinely pass the Turing Test, a measure of how well a machine can pretend to be a human.

2045 The technological singularity will occur as machines surpass people as the smartest life forms and the dominant species on the planet and perhaps Universe.

If Kurzweil is correct about these long-term predictions, as he was correct so many times in the past, it would raise new and sinister issues related to our future in the age of intelligent machines. About 10,000 scientists around the world work on different aspects of creating intelligent machines, with the main goal of making such machines as capable as possible. With amazing progress made in the field of AI over the last decade, it is more important than ever to make sure that the technology we are developing has a beneficial impact on humanity. With the appearance of robotic financial advisors, self-driving cars and personal digital assistants, come many unresolved problems. We have already experienced market crushes caused by intelligent trading software, accidents caused by self-driving cars and embarrassment from chat-bots which turned racist and engaged in hate speech. We predict that both the frequency and seriousness of such events will steadily increase as AIs become more capable. The failures of today's narrow domain AIs are just a warning: once we develop general artificial intelligence capable of cross-domain performance, hurt feelings will be the least of our concerns.

Methods

My discussion so far may seem to foreshadow impending instability because of AI. Popular discussion on AI and algorithms tends to share a similar tone. I proposed to try to cut through the hype with analytic and cross-disciplinary thinking on the risks and future of AI.

I convened a team of 12 researchers from across the academic disciplines and with a multitude of professional experiences to discuss AI. We curated a team of colleagues who were diverse in gender, ethnicity, and race while also making sure that we did not over represent for deep technical knowledge of AI. The team included expertise in information technology, psychology, political science, engineering, mathematics, radiology and design. Our hope was that, by convening such a group of researchers with extensive and varied training, we would encourage a dialogue around AI that was distinct and would allow for insights from topics and substance adjacent to AI.

The group's first exercise was to take part in a structured brainstorming session involving independent thought, small group discussions, and whole-group debate to first develop a working definition of AI and then to highlight application areas most prone to disruption by AI. The working definitions were developed via a rapid-fire collection of answers to "describe AI in less than five words." The themes of the contributions were summarized.

We followed these initial exercises by driving the team to deeper discussion via a future-casting exercise for which the larger group was split into 4 subgroups.

Results

While the insights and outcomes of the activity varied depending on the envisioned future that the groups chose, there was a list of application spaces that were included by each of the teams.

Due to the consistency across groups, I consider those to be "no-brainer" applications of AI, and they include

Cyber-security
Security (national and domestic)
Employment ("future of work")
Decision making
Health.

Following the activities with our team of colleagues, I chose to dive more deeply into the literature on the first topic; Cyber-security with a goal of developing a clearer picture of the risks inherent in the use of algorithms or artificial intelligence (or jointly as artificial agents) in these spaces. We chose these topics because we believe they are more pressing concerns to governments and the populace. To discuss AI benefits and risks to cyber-security and the role of humans as related to the security of nation states, we convened a smaller team of colleagues with deep knowledge of systems, algorithms, programming and security.

Discussions

First, the cyber environment is becoming more and more complex along with the cyber-threats. For example, "By 2020 Cisco estimates that 99% of devices (50 billion) will be connected to the Internet. In contrast, currently only around 1% is connected today" (e.g., Rosenquist, 2014). Even defenses are becoming complex, whether a defense is passive or active (e.g., despite our lengthy review of cyber defenses, we omitted numerous defenses, such as the use of encrypting emails, randomly generating passwords, using peer networks to increase security, hardening websites, etc.; from FIPSP, 1993; Intel, 2014; respectively). One of the problems with defending a website against cyber threats is that the relative value of what is being protected increases to cyber-attackers as the defenses they face improve, fueling the arms race between cyber hackers and cyber defenders (Schwartz, 2014).

This review was not inclusive of all potential cyber threats. We omitted many threats, such as those for businesses that must handle private personnel information.

With advancement, new exploits and vulnerabilities could be easily identified and analyzed to prevent further attacks. Incident response systems could also benefit greatly from AI. When under attack, the system will be able to identify the entry point and stop the attack as well as patch the vulnerability.

Studies show that it takes, on an average in 2016, 99 days for a company to realize that they have been compromised. Although a long way from 146 days in 2015, yet a very long time for the attackers to gain all

DOI: 10.21522/TIJAR.2014.SE.19.01.Art001
ISSN: 2520-3088

the information they were looking for. This time frame is not only enough to steal data but also manipulate it without detection. This can have a great impact on the company as it makes it very difficult for the company to differentiate between the fake and the actual data.

But, in addition, we want to understand how malicious agents interdependently select targets – not just watch them do it. We should be able to create a system that predicts a malicious action before a red team composed of humans or autonomous AI agents enact a threat. Based on data sets of past cyber threats and defensive actions, predictive cyber threat analytics that predict future threats should become a part of the AI tool kit used by defenders against malicious actors.

From an individual perspective, cognitive biases form individual vulnerabilities that cyber-attackers attempt to exploit.

Advantages of artificial intelligence

Organizations face millions of threats each day making is impossible for security researcher to analyze and categorize them. This task can be done by using Machine Learning in an efficient way.

By finding a way to work towards unsupervised and supervised machine learning will enable us to fully utilize our current knowledge of threats and vectors. Once those are combined with the ability to detect new attacks and discover new vulnerabilities, our systems will be able to protect us from threats is a much better and efficient way.

However, like every Machine Learning algorithm, even these advanced algorithms would require human guidance to learn from as humans are better equipped to look beyond a simple anomaly that a machine could pick up and put it in a different context and decide to ignore it as a security threat.

Another benefit of using AI based machines is that, in theory, these systems would work in a more calculated approach and in a more accurate way resulting in eliminating human error. Additionally, these systems could work simultaneously on various tasks, monitoring and protecting a vast number of devices and systems. They can therefore mitigate large scale attacks.

Disadvantages of artificial intelligence

The biggest disadvantage of any AI based system is that we cannot predict what it'll do. If fallen into the wrong hands, the result could be fatal and a whole different level can could do more damage than good.

A super-intelligent AI will be really good at completing goals, but, if those goals aren't aligned with ours, we'll have a problem. AI in security systems had foregone the utilization of valuable analyst skills and therefore didn't benefit from human feedback.

Even though the initial concerns about the development on AI in cyber security may revolve around concerns about eliminating the much-needed human expertise, intuition and judgment, the real disadvantage of artificial intelligence is its unpredictability.

Conclusion

We first agreed that cyber threats are making cyber environments more complex and uncomfortable for average users; second, we concluded that various factors are important (e.g., timely actions are often necessary in cyber space to counter the threats of the attacks that commonly occur at internet speeds, but also the 'slow and low' advanced persistent threats (APTs) attacks that are difficult to detect, threats that occur only after pre-specified conditions have been satisfied that trigger an unsuspecting attack). Third, we concluded that APTs pose a risk to users but also to national security (viz., the persistent threats posed by other Nations). Fourth, we contend that using "red" teams to search cyber defenses for vulnerabilities encourages users and organizations to better defend themselves. Fifth, the current state of theory leaves many questions unanswered that researchers must pursue to mitigate or neutralize present and future threats. Lastly, we agree with the literature that cyber space has had a dramatic impact on human life and that the cyber domain is a breeding ground for disorder. However, we also believe that actions by humans and AI researchers can be taken to stay safe and ahead of existing and future threats.

Fully autonomous machines can never be assumed to be safe. The difficulty of the problem is not that one particular step on the road to friendly AI is hard and once we solve it, we are done. All of the steps on the path are simply impossible. First, human values are inconsistent and dynamic and so cannot be understood and subsequently programmed into a machine. Suggestions for overcoming this obstacle require changing humanity into something it is not, and so by definition destroying it. Second, even if we did have a consistent and static set of values to implement, we would have no way of knowing if a self-modifying, self-improving, continuously learning intelligence greater than ours will continue to subscribe to that set of values. Perhaps, friendly AI research is exactly what will teach us how to do that, but we think fundamental limits on verifiability will prevent any such proof. At best we will arrive at a probabilistic proof that a system is consistent with some set of fixed constraints, but it is far from "safe" for an unrestricted set of inputs. Additionally, all programs have bugs, and can be hacked, or malfunction because of natural or externally caused hardware failure, etc. To summarize, at best we will end up with a probabilistically safe system.

It is really difficult to predict what the future Artificial Intelligence holds. Some say it'll help us to better our world while the others lean towards the possibility that it may go rouge.

References

[1].Acemoglu, Daron, and Pascual Restrepo, "Robots and Jobs: Evidence from US Labor Markets," National Bureau of Economic Research, NBER Working Paper No. 23285, 2017. As of October 11, 2017: http://www.nber.org/papers/w23285

[2].Anderson, James M., Nidhi Kalra, Karlyn Stanley, Paul Sorensen, Constantine Samaras, and Tobi A. Oluwatola, Autonomous Vehicle Technology: A Guide for Policymakers, Santa Monica, Calif.: RAND Corporation, RR-443-2-RC, 2016. As of October 11, 2017: https://www.rand.org/pubs/research_reports/RR443-2.html

[3].Artificial Intelligence Safety and Cybersecurity: A Timeline of AI Failures by Roman V. Yampolskiy and M. S. Spellchecker from https://arxiv.org/pdf/1610.07997.pdf

[4].Artificial Intelligence and its impact on Cyber Security from https://medium.com/@chiraghdewan/artificial-intelligence-and-its-impact-on-cyber-security-1b2446d770b9

[5].Autor, David H., David Dorn, Lawrence F. Katz, Christina Patterson, and John Van Reenen, "The Fall of the Labor Share and the Rise of Superstar Firms," CEPR Discussion Paper No. DP12041, May 2017b. As of October 11, 2017: https://ssrn.com/abstract=2968382

[6].Axelrod, R. (1984). The evolution of cooperation. New York, Basic.

[7].Baiocchi, Dave, and D. Steven Fox, Surprise! From CEOs to Navy SEALs: How a Select Group of Professionals Prepare for and Respond to the Unexpected, Santa Monica, Calif.: RAND Corporation, RR-341-NRO, 2013. As of November 16, 2016: http://www.rand.org/pubs/research_reports/RR341.html

[8].Baker, Brian J., "The Laboring Labor Share of Income: The 'Miracle' Ends," Monthly Labor Review, U.S. Bureau of Labor Statistics, 2016. As of November 16, 2016: http://www.bls.gov/opub/mlr/2016/beyond-bls/the-laboring-labor-share-of-income- the-miracle-ends.htm

[9].Barocas, S., and A. D. Selbst, "Big Data's Disparate Impact," California Law Review, Vol. 104, 2016.

[10]. Bayern, Shawn J., "The Implications of Modern Business-Entity Law for the Regulation of Autonomous Systems," Stanford Technology Law Review, Vol. 19, No. 93, October 31, 2015. As of October 11, 2017.

[11]. Cieply, M. & Barnes, B. (2014, 12/30), "Sony Cyberattack, first a Nuisance, Swiftly Grew into a Firestorm", New York Times, from http://www.nytimes.com/2014/12/31/business/media/sony- attack-first-a-nuisance-swiftly-grew-into-a-firestorm-.html

[12]. Daniel Merkle; Martin Middendorf (2013). "Swarm Intelligence". In Burke, Edmund K.; Kendall, Graham. *Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques*. Springer Science & Business Media. ISBN 9781461469407.

[13]. Definition of AI as the study of intelligent agents: Poole, Mackworth & Goebel 1998, p. 1, which provides the version that is used in this article. Note that they use the term "computational intelligence" as a synonym for artificial

DOI: 10.21522/TIJAR.2014.SE.19.01.Art001
ISSN: 2520-3088

intelligence. Russell & Norvig (2003) (who prefer the term "rational agent") and write "The whole-agent view is now widely accepted in the field" (Russell & Norvig 2003, p. 55). Nilsson 1998, Legg & Hutter 2007.
[14]. F. Pistono and R. V. Yampolskiy, "Unethical Research: How to Create a Malevolent Artificial Intelligence," presented at the 25th International Joint Conference on Artificial Intelligence (IJCAI-16). Ethics for Artificial Intelligence Workshop (AI-Ethics-2016), New York, NY, July 9, 2016.
[15]. Human Factors in Cybersecurity and the Role for AI by Ranjeev Mittu & William F. Lawless from https://www.aaai.org/ocs/index.php/SSS/SSS15/paper/viewFile/10248/10054
[16]. Lawless, W.F., Mittu, Ranjeev, Marble, Julie, Coyne, Joseph, Abramson, Myriam, Sibley, Ciara & Gu, Wei (2015, forthcoming), The Human Factor in Cybersecurity: Robust & Intelligent Defense. To be published by Springer.
[17]. LII (2014), 44 U.S. Code § 3544 - Federal agency responsibilities, Title 44, Chapter 35, Subchapter III, Legal Information Institute at Cornell University Legal School, from 44 USC §3542; see http://www.law.cornell.edu/uscode/text/44/3544
[18]. Loukas, G., Gan, D. & Vuong, T. (2013, 3/22), A taxonomy of cyber-attack and defence mechanisms for emergency management, 2013, Third International Workshop on Pervasive Networks for Emergency Management, IEEE, San Diego.
[19]. Maloof, Mark. "Artificial Intelligence: An Introduction, p. 37" (PDF). *georgetown.edu*.
[20]. Martinez, D., Lincoln Laboratory, Massachusetts Institute of Technology (2014, invited presentation), Architecture for Machine Learning Techniques to Enable Augmented Cognition in the Context of Decision Support Systems. Invited paper for presentation at HCI.
[21]. The Risks of Artificial Intelligence to Security and the Future of Work Osonde A. Osoba, William Welser IV from https://www.rand.org/content/dam/rand/pubs/perspectives/PE200/PE237/RAND_PE237.pdf
[22]. R. V. Yampolskiy, "Taxonomy of Pathways to Dangerous Artificial Intelligence," in Workshops at the Thirtieth AAAI Conference on Artificial Intelligence, 2016.